SDS PODCAST EPISODE 795: FAST-EVOLVING DATA AND AI REGULATORY FRAMEWORKS, WITH DR. GINA GUILLAUME-JOSEPH

Jon: 00:00:00 This is episode number 795 with Dr. Gina Guillaume-Joseph. Today's episode is brought to you by AWS Cloud Computing Service, by Crawlbase, the ultimate data crawling platform, and by Babbel, the science-backed language-learning platform.

00:00:20 Welcome to the Super Data Science podcast, the most listened to podcast in the data science industry. Each week, we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now let's make the complex simple.

00:00:51 Welcome back to the Super Data Science podcast. We've got an excellent episode for you today with Dr. Gina Guillaume-Joseph. Gina was until recently the CTO responsible for government at Workday, aligning that HR tech giant with the US federal government's tech transformation strategy. Prior to Workday, she was director of technology at the financial giant, Capital One. Earlier, she spent 16 years supporting the federal government as a contractor with leading firms like Booz Allen Hamilton and the MITRE Corporation. She now works as a fractional Chief Information Officer and as a Adjunct Faculty at George Washington University. She holds a PhD in Systems Engineering from George Washington University and a Bachelor's in Computer Science from Boston College.

00:01:33 Today's episode should be of interest to just about anyone who would listen to this podcast because it focuses on the data and AI regulatory frameworks that will transform our industry in the coming years. In today's episode, Gina details the dark data conundrum, the most important data and AI regulations of recent years, as well as those that are coming soon, the pros and cons of being or hiring a fractional executive and what systems engineering is and why it's an invaluable background for implementing

		large scale AI projects. Are you ready for this important episode? Let's go.
	00:02:12	Gina, welcome to the Super Data Science podcast. It is my delight to have you here on the show today. Where are you calling in from today, Gina?
Gina:	00:02:20	I'm calling from Leesburg, Virginia.
Jon:	00:02:23	Nice. That's kind of in the Washington, DC metro area kind of thing?
Gina:	00:02:27	That's correct.
Jon:	00:02:28	Nice. Yeah, I think pretty much everyone who works in, there's very few people who live and work in DC. Isn't that right? It's kind of like a commuter hub around there.
Gina:	00:02:38	It is, yeah. We come from Maryland, Virginia to work in DC, yes.
Jon:	00:02:44	Mm-hmm. And famous for its traffic as a result.
Gina:	00:02:47	Wooh! Yes. You don't want to be caught up there during rush hour.
Jon:	00:02:54	So we met each other in New York, another place where you are very likely to be crawling along in your vehicle. We met at the Data Universe Speaker's Dinner, which was an exquisite meal and a very well organized event overall. Looking forward to seeing it. It was just the first ever Data Universe conference. It was held at the Javits Center, which is huge in New York City. And years ago, pre-pandemic, that's where O'Reilly used to host their big New York conferences like Strata + Hadoop World. And so hopefully Data Universe can start to grow and grow over the years and take that place and be kind of this huge must-not miss conference in New York City.

00:03:39 While the conference is running, you were speaking about bridging the divide. So tech's role in shaping policy for a sustainable future, and so that caught my attention immediately as a topic that would be fascinating to have on air because we seldom on the show, it's been years at least since on this podcast, we've focused on the interplay between technology and policy.

00:04:09 So to start things off, you introduce this idea of the dark data conundrum. And so you provided some estimates that there will be 175 zettabytes of data globally, which is basically just a huge number. There's 21 zeros behind the 175. And the key point here isn't so much the size, which is unfathomably large, but the idea is that only a very small percent of those data are used and analyzed. Is that right?

Gina: 00:04:46 That is absolutely correct, yes. Oh, so 175 is 21 zeros, and 80% of that data is unstructured, meaning that it was used one time for a data decision. It was an image, a picture file, a data file, a Google Drive, an email that was used once time and then it was stored away. And then 90% of that unstructured data is never analyzed, meaning that it was used one time, but never delve deeply into a full data set. And so that's that dark data conundrum because it's sitting out there maybe protected, but also maybe vulnerable, and then not leveraging the full use case and the full capabilities of that data to really make good structured, consolidated decisions that that is.

Jon: 00:05:43 80% unstructured, like you're saying, probably never used for anything else. Of those 80% that are unstructured, you had another stat in your talk that 90% of those are never analyzed in any way. And that huge amount of data, the vast majority in fact of data that you call these dark data that are unused, unstructured, those could be powerful fodder for AI systems. We talk about

situations where these large language models like GPT-4 are trained on all of the structured information available on the internet, all the kind of text that Google can scrape or OpenAI can scrape. And so this huge amount of data, these dark data, could provide a lot more fodder for AI systems could potentially be put to good use.

Gina: 00:06:41 Yes, it absolutely can. And figuring out how to do that I think is where we're coming to in terms of this AI age, this machine learning age. But it also comes with some challenges as well. When you have such large data sets, you have regulatory compliance issues because not all data is accessible. Or you can access it, there are regulations around it, and navigating those regulations can be a challenge. There's some privacy and security concerns around the data, and that's sometimes why it's not used. And so safeguarding human data, personally identifiable data, health data, all of that is really important.

00:07:25 And then when you're building the algorithms, when you're building the AI models, you have to ensure there's algorithmic fairness against bias. There might be some economic and workforce impacts, managing impact on employment. There's technical limitations. There's so many challenges around trying to understand and bring the data together and gain insights from it. And then I'm going to talk about the environmental impacts. When you're building these large data centers, how does that impact the environment? How does it impact earth and the people around us?

Jon: 00:08:02 Nice. Yeah, it makes perfect sense. Lots of things to consider here when trying to capitalize on all of these dark data. So let's start off with, I guess, what responsible AI means and how we can have a framework for addressing the kinds of concerns that you just described. So ethical considerations, privacy concerns, biases in the

algorithms that eventually show up, what is the need? Well, I mean, you've actually kind of already outlined the need, but how do we do it? Like, what, in your view, should governments ideally be doing? And we're going to go into then some detail on what... Particularly in the US where you have expertise, we're going dig into specific examples of current or prospective legislation that is designed to protect data or regulate AI systems. So we'll get into those specific examples afterward, which kind of in general, what in your view makes an AI framework comprehensive or useful?

Gina: 00:09:13 The robustness of an AI policy and the governance around it, it makes it a responsible AI framework because the data is everywhere. You've got air traffic control data, you've got factory data, you've got technical data, you've got data in your homes. I've got a Nest system, and I've also got one of those keyless entry devices on my front door. And so that's data of the entry of my kids, of our face, biometrics, all of that. And the need to manage it and to govern the data is critical because, again, of the proliferation of this large unstructured swath of data that's really sitting in these data centers.

00:10:07 So being able to create that responsible AI framework use, eliminating those challenges, the biases, ensuring security, ensuring privacy, ensuring protection, all of that is a critical aspect of it. And that's why early 2023, there was the executive order, the president's executive order on AI. What they were trying to do is wrangle it, put some guardrails around manipulating the data, accessing the data, understanding the data, and then the usefulness of the data using it in a way that is responsible and that is protecting humans that essentially are the ones that's pulling the data and having access to the data. That's the critical part of what needs to happen.

00:10:56 And we, because it's so early, we don't know what we don't know. There's so many unknown unknowns that we have to come together collectively to be able to help the federal government develop those policies, help the NIST and MITRE and others to come together to figure out what does that look like? What does a responsible AI framework look like? I don't personally know, but if we all come together and collectively bring our insights together from across industries, we can be able to build that out.

Jon: 00:11:34 Yeah. And you say you don't personally know, but this is something that you are expert in. I guess the point there is that no individual can themselves grasp all of the different kinds of expertise specializations that are required to have successful AI frameworks, AI policy. So that all makes perfect sense.

00:12:01 And so maybe this is kind of a question that just came to me off the top of my head. I know that we're going to be mostly in this episode talking about the US federal government and its role in AI regulation. Before we dig into that, a kind of meta question that sits on top, I suppose in a way, is why should it be the federal government say, as opposed to states or local regulators that are leading the charge here? Do you think that there's advantages? Or what are the pros and cons of it being the federal government? Why is that going to be our focus today?

Gina: 00:12:38 Well, they start the conversation. It's the federal government. It's their mission. We have these agencies that are set up that are mission-focused to solve problems for the people of the United States, right? You've got the VA, Social Security Administration. They were all enacted by Congress to solve a problem. And we know that there's some challenges with AI, and so they're in the best place to be able to set up and stand up organizations, to stand up and develop these regulations and policies with the

support of private industry and academia to be able to manage and guide ethical AI development and use, but also at the same time, balancing innovation, transparency, accountability, and the public interest because the federal government works for the public interest. They're the ones that are taking the lead.

00:13:36 And we see that with the executive order on AI as well as there's the AI Bill of Rights that lays out the five components for ethical, safe, trustworthy, privacy, explainability, and then the human considerations in the fallback for AI systems. And so all of that coming together with private industry will help to build a robust framework, so to speak, because it's private industry that has specific regulatory compliances that they must also abide by in order to ensure that the data is safe, that it's private, there's privacy concerns, their security concerns are all banking institutions. I work for Capital One. Again, it's a concerted effort, but federal oftentimes leads the way when there's these big changes in society.

Jon: 00:14:39 This episode of Super Data Science is brought to you by AWS Trainium and Inferentia, the ideal accelerators for generative AI. AWS Trainium and Inferentia chips are purpose by AWS to train and deploy large scale models. Whether you are building with large language models or latent diffusion models, you no longer have to choose between optimizing performance or lowering costs. Learn more about how you can save up to 50% on training costs and up to 40% on inference costs with these high performance accelerators. We have all the links for getting started right away in the show notes. Awesome. Now back to our show.

00:15:19 Makes a lot of sense, everything that you just said. I suspect also in addition, of course, the federal government is protecting the largest number of people. So that's also... It's going to have the biggest impact. It's

typically going to have a bigger budget than a state would for enacting this kind of legislation. And also, I imagine from private company's perspectives or even universities' perspectives, if we can have the federal government come up with legislation that generally states and smaller regions feel like this has hit the nail on the head, then we don't need a patchwork of different regulations that different companies, universities, organizations, NGOs, all need to be going through hoops to be like, "Okay, if we want to operate in California and in Massachusetts, we're going to have to do these completely different things and store data in completely different ways." So yeah, I imagine it's also something that is in the favor of us as consumers and as people working in private companies having a legislation be considered at that federal level.

Gina: 00:16:27 Absolutely. I created a timeline for the talk in New York, but in 2019 there was the executive order on maintaining AI systems. And then in 2020 there was OMB guidance on AI regulation. There was the National AI Initiative Act of 2020. And then there's the federal acquisition regulation rule on AI. There's the AI in Government Act. There are a number and continuingly ongoing that attempt to bring some structure around the data that's used to create these models. And now what the federal government is doing is saying, "Hey, we can't do this on our own." And so they've set up these industry-focused groups that are also coming together, built out of a number of agencies, private industries, who are also sitting at the table with the federal government providing guidance and insights into what they feel should be some of those regulations around AI and machine learning.

Jon: 00:17:44 Nice. Yeah, thank you for giving that overview. And maybe if it's okay with you, we can actually include the slides from your talk as something in the show notes here, and so people will be able to see you put together this great timeline in that deck of these various executive orders

and regulations on AI that have been put in place. In a very short period of time, I mean, you highlight eight different categories of legislation that have been put together just in the last five years. So fast moving space.

Gina: 00:18:16 It's a fast moving space. And I think this year and last was when, I mean, Pandora's Box with ChatGPT.

Jon: 00:18:25 Mm-hmm. Exactly.

Gina: 00:18:27 Yeah, and all the movement there.

Jon: 00:18:29 Nice. And so we will dig into some of these specific executive orders and regulations in the US in a moment. We'll kind of highlight the most important ones, their impact on us as people who develop AI systems or use AI systems. But before we get to that, another thought that has occurred to me is that I kind of have this perception of the US federal government with respect to data privacy and AI, maybe digital rights in general. Maybe even rights in general, human rights in general. It seems like the US is often a step or two behind, say, the EU. And so then we end up in situations where EU legislation kind of forces tech companies, which are disproportionately US-based, all of the world's biggest tech companies are headquartered, at least in terms of their people, maybe not for tax purposes, but in terms of people they're headquartered in the Bay Area really of the United States.

00:19:36 And so it's interesting that we have this situation over time where historically at least it's the EU, and now even recently with AI, the EU AI Act seems to have been cemented before as sweeping AI legislation has been enacted in the US. And so that means that these US-based tech companies, these big tech companies are already trying to comply probably with the EU AI Act. And so I don't exactly have a question here, but I think you

probably have something interesting to add onto here to add some more color to this situation.

Gina: 00:20:18 What you just said is that AI is global, and we know that. We know that with the large swath of data that's unstructured, it's sitting at data centers across the globe. Google's got data centers everywhere across the globe. I live in Loudoun County, Virginia. It's the data center capital of the world. 75% of the world's internet traffic flows through Loudoun County. So what that means is that we have to come together as a collective front, a concerted effort globally to create these policies, regulations, to protect the data.

00:21:02 And maybe the EU is ahead, but we all must play our part. Every country, every municipality has to really come together and collaborate on creating these regulations that protect the data from privacy, protect the data from threat actors, to protect the data, health data, banking data, institution data. It's a collective effort. And I don't know if one's ahead and one's not, but I think we just have to share across county lines, country lines, data lines, to be able to really create a robust framework to help manage this growing explosive thing that's AI and machine learning. And fostering collaboration, fostering an engaging environment, to be able to do that, I think is critical for us to do.

Jon: 00:22:18 Nice. All right. That was a great explanation. Absolutely makes perfect sense that it's something we need coordination. And I think we have even seen instances, you may be able to remember this specific event better than I can, but there have been things like not just in the White House, but also in 10 Downing Street in the UK, we have private leaders like Sam Altman from OpenAI and other tech leaders meeting with government leaders and multiple governments, multiple federal governments being

involved from say the UK and the US altogether, and trying to shape these things together.

00:23:03 It seems to make the most sense in the same way with what I was describing earlier, where having federal legislation means that we don't have this patchwork of individual state legislations that companies need to be working around. In the same way, if we can have frameworks that are relatively homogenous across the globe, there's this situation that can happen where you could theoretically end up, if you have lots of complex regulation and lots of different jurisdictions, you can end up actually entrenching the incumbent big tech companies because they're the only ones that can have large enough legal teams and policy teams in order to be able to follow along with and adapt to the AI legislation in all these different regions.

00:23:57 So you lock in the big tech players and you make it more difficult for upstarts/startups to disrupt if you had all that kind of thing. So if in theoretically we could kind of have this global/universal AI Bill of Rights and you just had one set of policies that every company knew that they needed to adhere to, that's probably fanciful that we could get into that kind of scenario. But it would be an ideal because it would mean that startups, you would just have to understand one set of rules and they'd be able to compete more easily with the big tech players.

Gina: 00:24:31 And we see the proliferation of AI startups. They're really working to try to solve some of these challenges that we talked about early on with AI. And if we encumber them with these hard, fast rules and regulations, you're going to stifle creativity and innovation. And honestly, it's some of the smaller companies that really are the backbone of global economy, right? And so making sure that we come together collectively to create this policy, but it's a high level policy and each sector then develops their own

specific regulatory requirements to make sure that they're doing everything that they can to safeguard the data. And then bringing in the smaller AI startup companies as well to fill in some of those gaps that they might not be able to do because they're so large and they're so encumbered.

Jon: 00:25:34 Nicely said. Yeah. All right, so now that I've distracted long enough by expanding from state regulation to federal and then going globally, let's dig in now to specific federal leg islation. So we talked about how you'd identify these eight categories of AI regulations that have happened over the last five years. So let's dig into the AI Bill of Rights first, which is the one that I've heard talked about the most in the US in recent months.

00:26:05 And so you're the expert on this, but my understanding is that this comes from the White House Office of Science and Tech policy. So tell us about the AI Bill of Rights. And also I guess something that would be helpful here, maybe even the place to start, is what is the difference between something that is an executive order, like I believe this is, the AI Bill of Rights, and other I guess legislation that goes through Congress. What is the difference there? Maybe that's actually a good place to start before digging into the AI Bill of Rights itself.

Gina: 00:26:40 The AI Bill of Rights outlines how to build safe and effective systems, how to protect against algorithmic discrimination, bias and protection. What one needs to do in terms of making sure that data privacy is incorporated? Explainable AI, and AI that puts humans in the loop. It kind of says, "Here's what you should do." And then Congress might come in. Currently, I think there's 150 bills being worked on and private industry and other organizations are in there-

Jon: 00:27:21 150 bills just in general.

Gina: 00:27:22 In general.

Jon: 00:27:24 Yeah. So there might only be a few that are related to AI, but Congress is working on 150 different bills-

Gina: 00:27:29 Different bills.

Jon: 00:27:30 Potentially enshrining them into law consider... You know?

Gina: 00:27:33 Yes.

Jon: 00:27:34 Yeah, yeah, yeah.

Gina: 00:27:36 Yes. And they want to enshrine about, I think five to 10 into law this year in 2024 and through 2025, that they've identified with the support of private industry academia, the ones that they'd like to put into law, put into practice, part of regulatory compliances, part of banking legislation, part of tech legislation, institutional data protection. That's where you got to make sure when you're building out the systems that you've checked the boxes for these data privacy concerns, these NIST notifications, these safe and effective systems as you're building because then you've got a regulatory compliance.

00:28:22 You have auditors that come in to really check and make sure that your systems are in compliance. And that's where that flow down comes from. You get the AI Bill of Rights and then you've got congressional mandates, then you've got regulations that are coming through. We don't have a lot right now because AI is so new, and that's why we all have to come together to really help and support developing and building out the right set of regulations so that you don't stifle innovation, you don't crush the smaller AI startups and you allow creativity and innovation to really flourish.

Jon: 00:29:09 Hello, Super Data Science podcast listeners. Today's episode is brought to you by Crawlbase, the premier data crawling and scraping platform designed for anyone needing reliable data. Forget about hardware, infrastructure, proxies, set up blocks and CAPTCHAs. With Crawlbase, they handle all of that for you. Simply call the Crawlbase API and gather website data. They literally support millions of different websites. It's super easy. Try out Crawlbase today. As a offer to Super Data Science podcast listeners, use our exclusive code SuperDataScience, with no spaces, to unlock 10,000 free requests of value of $42. You'll also find the code in the podcast description, head over to Crawlbased.com and start crawling in minutes.

00:29:55 Yeah. And getting that, it is critical that we strike the right balance in order to facilitate exactly what you're describing there. You don't want to stifle innovation. The EU, while potentially in some digital privacy, in some AI respects, seems to, in my relatively uneducated view, to be somewhat ahead of the game relative to the US that has also been associated with stifling innovation. I mean, there are not household name tech companies based out of the US. Or sorry, based out of the EU. So they're completely wrong. They are mostly based out of the US and whereas the EU has not had many well-known international tech companies, and there could be all kinds of other cultural factors there.

00:30:48 But the thing that people point to more than anything else that I see out there is having privacy regulations, for example, that are too stringent and potentially in the future AI regulations that are too stringent that lock down that innovation. But yeah, AI Bill of Rights. So basically the AI Bill of Rights, that doesn't, today have... Because it sounded there like the AI Bill of Rights came out, then Congress kind of figures out what to actually enshrine into law based on potentially the AI Bill of Rights. And

then we get downstream from that. We get regulations that are available to us to follow.

00:31:31 So with the AI Bill of Rights, as it stands today, it's a guiding light. It isn't actually something that we can go and look up today and say, "Based on this AI Bill of Rights, I know as a company what I can do, or I know what rights I have with AI as a consumer." It's just something that we can look up today and maybe it gives us a sense of if we're developing a technology, we can say, "Okay, based on this blueprint for an AI Bill of Rights, I can expect that in a year or two, legislation that looks like this or regulations that look like this might be in place, and so therefore all want to be careful about how I collect data today." That's kind of how I can use the Bill of Rights today. Is that right?

Gina: 00:32:18 Yes. So companies take the responsibility to really build these systems right, to build them in a way that protects the human, that protects the technology, protects the data. And then you bring in the National Institute of Standards. They've also introduced the AI risk management framework where they provide recommendations, they provide the requirements and say, "Here's how to design, develop, use, and evaluate AI systems." You want to create multidisciplinary teams to address risks throughout the AI systems lifecycle. The risks that we talked about earlier in the podcast, ensuring inclusive design, responsible development, ethical use, continuous evaluation. Because AI is moving so fast, you got to continuously monitor in order to ensure that it's remains safe and secure.

00:33:20 And then there's also the aspect of, because there aren't any regulation, voluntary adoption. You encourage organizations to adopt the framework, demonstrating its benefits for risk management, for building trustworthy AI, providing resources and supporting the implementation.

There's also that aspect of collaboration and consensus that we talked about. Again, because AI and these systems and these data centers live all over the world, you want to foster that collaborative environment to ensure that the data is secure, to ensure that you're engaging all the stakeholders from various sectors in the discussion and decision-making processing because you want to leverage the collective expertise to refine the framework that we want to build out.

00:34:17 And then managing risk to the individuals, to the organizations and to society. Remember, AI cross country lines, so it's societal impact can be large. So that methodology is important, that identifies, that evaluates, that mitigates all these potential harms to society organizations and the individuals around, again, privacy security and ethical considerations. And then alignment with global AI risk management initiatives. Again, that global perspective is really critical. And then it drives that open, consensus-driven, transparent and collaborative development process that we want to get into the mindset of doing.

Jon: 00:35:07 Nice. Great. So I think I'm starting to piece all these different puzzle pieces together. So explain for us, you mentioned there, so there's the NIST AI Risk Management framework. So NIST is the National Institutes of Science and Technology. That may be familiar as an acronym. The NIST thing is something that for those of us who have done any deep learning, the kind of "Hello world" deep learning example involves this handwritten data set of digits.

00:35:39 So it's 60,000 handwritten digits done by, if I remember correctly, US postal workers as well as elementary school students. And so it's just each image is a different digit. So some of them are zero, some are one, some are two, some are threes all the way up to nine. This handwritten

dataset was curated initially, I guess in the '90s, maybe even in the '80s by NIST. And then Yann LeCun, who's one of the most famous AI researchers of all time, he modified with his research team, at the time I believe they were AT&T Bell-Labs, they modified that NIST handwritten digit dataset to create the MNIST, modified NIST handwritten set. So I don't know, it's a bit of an aside, but that MNIST dataset is probably familiar to anyone who's done any kind of deep learning at all.

00:36:32 And so yeah, so that same organization, NIST, has been around for a long time in the US, I don't know how many decades, but has been trying to set up frameworks for all different kinds of industries in science and technology and has now created this AI risk management framework, which again, I'll have a link to that in the show notes alongside the AI Bill of Rights.

00:36:54 A third framework, I guess, you can correct me if I'm not using the right word there, that you brought up in your talk that also seems really helpful here is something called the MITRE ATLAS. So I've been trying to, as you've been speaking, kind of dig up what MITRE stands for, M-I-T-R-E. It doesn't seem like it stands for anything. Can you tell us a bit about MITRE and the MITRE ATLAS? And then maybe you can weave together these three different things, the AI Bill of Rights, the NIST AI regulatory framework, as well as MITRE ATLAS, and tell us how we can integrate these three frameworks together in order to have a good sense of how to move forward with the AI systems that we built.

Gina: 00:37:39 So MITRE is a not-for-profit organization. I worked for them for 10 years, and they support the federal government across all the federal government agencies to help them solve some of their most pressing challenges. So MITRE operates federally funded research and development centers in support of the federal government

to solve problems for safer world essentially is what MITRE does.

00:38:09 And while at MITRE, I supported multiple agencies, department of Homeland Security, Social Security administration, the Veterans Affairs, Department of Defense in some of the challenges that they were facing at the time, societal challenges to include when the economy was doing some downward slides and things were failing. Part of some of the work that I did was at the FDIC, that was with Booz Allen, but MITRE was involved in other aspects of that as well, to really understand the failures and to figure out the mitigation strategies to ensure that society didn't feel those impacts as broadly and as strongly.

00:39:06 MITRE created the ATLAS Threat model introduction of Threat model. It's this comprehensive coverage of AI specific adversary tactics and techniques that includes real world observation and reporting. It talks about accessibility and usability of AI, alignment with existing cybersecurity frameworks in terms of from an AI perspective and that community engagement contribution and the educational resources and training. So they're developing a detailed taxonomy of tactics, of techniques, of procedures specific to AI systems that cover the entire life cycle from data collection to model development and deployment and maintenance, where they establish those mechanisms for continuously gathering and updating threat intelligence based on real world cybersecurity incidents involving AI so that the knowledge base remains current and relevant.

00:40:14 So that's what MITRE is doing it with their MITRE ATLAS framework. And the framework integrates their existing MITRE ATT&CK for enterprise framework that shows that they bring in that consistency and interoperability across

cybersecurity efforts as it pertains to AI systems. That's my ATLAS Threat model.

Jon: 00:40:42 Nice, nice, nice. So it sounds like in some ways similar to NIST, MITRE exists to keep people safe. And I guess NIST is maybe actually a bit more narrow in that it's specific to science and technology where it sounds like MITRE is, you gave examples there of financial safeguarding that it's involved with. So it sounds like MITRE is generally just trying to keep people safe. So how can we, getting back to this idea of how can we integrate the kinds of guidance that we get from NIST with MITRE, maybe even the AI Bill of Rights, how can we integrate that information in order to implement safe AI systems that are likely to be compliant with future regulations?

Gina: 00:41:33 So that's one of the things that I pulled together myself to figure out how do we bring them all together, like you said, the AI Bill of Rights, the NIST framework, MITRE ATLAS, and any of the congressional mandates that come out to really build out this holistic, robust set of guidance in this AI journey that we're all on now. And part of that is that holistic AI systems lifecycle approach. So looking across the lifecycle from initial design through development, through deployment and that ongoing maintenance, you want to make sure that you incorporate those specific adversary tactics and techniques at every stage to address any potential threats that are posed from leveraging the AI machine learning system and the data that it's being used.

00:42:27 Then there's that real world threat intelligence integration. Again, real world, we're not talking about the US, we're talking about the worldwide threat observation, developing those mechanisms that are continuously updating your threat intelligence, ensuring that AI risk management strategies are informed by the latest adversary tactics, relevant specifically to AI systems. And

then the framework accessibility and interoperability, again, across timelines, datelines and across country lines and boundaries, integrating that framework and making sure that it's accessible to that broad range of stakeholders, including AI developers, cybersecurity professionals, policy makers, and designing it to be interoperable with existing cybersecurity and AI standards, whether those come from EU, the United States, Singapore, China, or anywhere else, ensuring that you're developing it in a way that it's facilitating adoption across different sectors, industries, and countries. That's where I came out of how do we integrate and make this a one robust holistic framework.

Jon: 00:43:52 If you're a regular listener, you know that last year I did a European podcast tour interviewing incredible guests in Amsterdam, Paris, and Berlin. While all the guests spoke perfect English, Babbel was invaluable for me to learn and practice Dutch, French, and German, enabling me to get directions and order my meals in the local language. Super fun, rewarding, and in some cases, an essential skill. Now you can do the same with a special limited time deal right now. Get up to 60% off your Babbel subscription, but only for our listeners at babble.com/superdata. Get up to 60% off at babble.com/superdata, spelled B-A-B-B-E-L dot com slash SUPERDATA. Rules and restrictions may apply.

00:44:39 Fantastic, thank you. Yeah, and I will, again, I'll be including your slides with this, I guess, in your own personal proprietary blend of how we can be integrating NIST and MITRE ATLAS frameworks along with considerations like the AI Bill of Rights with specific concrete implementation recommendations based on your professional experience and keeping closely on top of all of these fast moving developments that are coming out of the US federal government.

00:45:14 So thank you so much for those insights, Gina. What do you think is next? So this is where you ended your talk at Data Universe. So what kind of opportunities are there as AI continues to evolve, what's coming later after this kind of first round of AI regulatory frameworks and legislation?

Gina: 00:45:38 I think part of that, because it's a fast moving train, you got to be dynamic, innovative, and pliable where you're constantly incorporating in any ongoing research, you're pulling in the dialogues across disciplines and sectors. You're adapting to these technological advances while also safeguarding the ethical principles and societal values, and that's that holistic approach to AI governance for a secure, ethical, technological future where you develop these adaptive regulatory frameworks, you promote international standards and cooperation.

00:46:19 You asked a great question. The EU's ahead and the US lags behind. If we all come together, we're going to move together at the same pace and at the pace of innovation, at the pace of this moving train that we call AI. That public-private partnerships, you got to bring private industry and public industries together to be able to develop this robust framework, that transparency and public engagement, making sure that the public is involved in developing some of these because their societies are the ones that are impacted.

00:47:03 Data centers, I think I talked about that a little bit. Data centers worldwide use 5.2 billion gallons of water each year. That was for 2022. What effect does it have on small communities, and how are they able to really mitigate the impacts when these small communities might have droughts, they might have challenges in ensuring safe, clean drinking water, and we're pouring it into the data centers. Some of them have revolted and blocked Google data centers from being built, which means that you got to really bring in the people. You got to bring in these

private citizens into this conversation so that you can mitigate AI's carbon footprint and its energy usage as well because it does impact the earth. It does impact from raw materials to build the data centers, the hardware, the energy infrastructure. We're talking about concrete, steel, aluminum, plastic. And then there's e-waste. How do you properly dispose of and manage and recycle these systems or these raw materials once they're used. That's all part of the conversation. That's all part of the conversation to establish ethical, safe AI systems that protect the humans and protect the earth.

Jon: 00:48:38 Nicely said. I'm glad that we were recording that. That's one of those things that I'm like, "Wow, that was super well delivered." I am glad we have that on film. Nicely done, Gina. Yeah, very nicely summarized. And yeah, it gives us hope that we can get this AI world which has potentially so many benefits across healthcare, agriculture, education. You name it, any industry can be transformed for the better with AI, but there are things like privacy, environmental concerns, ethical use, all these kinds of things that you've been talking about in the episode, including the environment stuff that's come up most recently that we need to get right in order for everyone to feel on this planet like AI is working for them, not against them. Very cool. So glad to have you doing that work.

00:49:34 Now, speaking of your work, if I understand correctly, you have done kind of fractional executive work for different companies where you split your time, your work week into fractions in order to bring this incredible knowledge that you have to a bunch of different companies simultaneously. What brought you into doing that kind of fractional work and who would you recommend it to or who would you not recommend it to?

Gina: 00:50:05 So the fractional work is, remember we talked about those small AI startups that want to really do good work, but they just don't have the budget because they're fundraising? Those are the small companies that I support in a fractional capacity aligned with their budget. And so I do the CTO work in terms of helping them understand what their roadmap should look like, what markets they should enter. Federal government is a long lead time in terms of getting access and getting into and selling, but creating that future roadmap to be able to do that in the future and to be able to raise funds to show the value of their product is some of the things that I help these smaller AI startups do on a fractional basis, splitting out my time across different ones.

Jon: 00:51:06 Yeah. There must be situations that you run into... I could imagine, this is off the top of my head, but it seems like in that kind of situation where you're splitting your time between different companies, company A and company B might both have something really urgent coming up today or this hour and they need you immediately for a meeting. Both companies need you at the same time or they need more of your time than you could possibly have in the same week. How do you manage situations like that as a fractional executive?

Gina: 00:51:51 I've never run against that actually.

Jon: 00:51:51 Okay. Yeah, yeah. So maybe you set up good-

Gina: 00:51:51 Good questions. I got to create a robust framework to be able to mitigate that risk.

Jon: 00:51:56 Yeah, maybe you set expectations well in advance. So then maybe that's the lesson here, is how do you set up expectations to make sure that everyone, all the firms that you're doing fractional work for are aware that you're doing that fractionally and that they understand that

you're not always going to be available at the drop of a hat because you might have other stakeholders that you need to work with?

Gina: 00:52:20 Yeah. So I do share calendars across. And I've got a robust set of color codings to be able to manage and know where I'm going to be. And I do podcasts like this and I do [inaudible 00:52:33] conferences. And so that's kind of what helps me in terms of managing. And it is not a bunch of companies, so we want to be clear. I work with maximum two at a time because I want to give them the best of me and the best of my knowledge so that I can bring them the best information at the right time when it drops. And so I do manage that, and I am cognizant of these AI companies. If they miss a contract deadline, it could be the end. And so I want to make sure that I give them all of me in the time allotted in the contract obligations.

Jon: 00:53:22 Nice. All right. Thank you for giving us that insight. It makes perfect sense. I guess it all is kind of common sense when you break it down like that. I do like the color coded calendar shared across those orgs, though. That's a helpful practical tip. Now, looking back a little further into your career, you obtained a PhD in Systems Engineering from George Washington University there in DC.

00:53:42 And so first of all, tell us what systems engineering is. I think we can pull up into our heads relatively easy. If somebody says chemical engineering or software engineering or material engineering, you can probably get a pretty good idea of like, "Okay, we're making software or making some kind of chemical or designing a chemical process to create a chemical," it seems relatively straightforward to imagine what the person is doing. What does a Systems Engineer do? What does that mean to be a Systems Engineer?

Gina: 00:54:14 So a Systems Engineer has the full lifecycle perspective and understanding. Remember when I talked about integrating the NIST AI framework and MITRE ATLAS. In order to manage and cover the practices from initial design through development, deployment, ongoing maintenance and disposal, that encompasses systems engineering, being able to work across that lifecycle. Incorporating security management, risk management, program management, all within that to ensure you're delivering products on time and on budget. That's the systems engineering lifecycle.

Jon: 00:54:52 Nice.

Gina: 00:54:53 And I got my PhD in systems engineering specifically, I developed a AI model in 2016 that looked at project failures and why projects failed and developed it with small dataset. We're not talking large language models here, 250 data points across the software development industry, looking at failed projects as well as successful project, created that predictive analytics model, leveraging R programming to be able to show why projects fail and the likelihood of occurrence of failure, and then developing the mitigation strategies to allow projects to really get back on track. And some of the major failure points was when projects go off schedule or they eat up their budget or they go off budget or insufficient knowledge and understanding of the product or the systems that's being built by the development team, stakeholder support, executive stakeholder support, or the lack thereof, some of those things cause projects to fail.

Jon: 00:56:11 Nice. And so that does give me a pretty good sense of what systems engineering is. And it's something that's always seemed very prestigious in my mind because I grew up... I went to high school at least in Waterloo, Ontario, Canada, which has the University of Waterloo,

which is one of the world's top engineering schools. And I remember when I was graduating high school, the program at the University of Waterloo that had the most difficult entry criteria, that was the most in demand was systems engineering. And so I have it on a pedestal in my head of being this extremely prestigious, challenging, and rewarding area of engineering.

00:56:54 Now since finishing your PhD about a decade ago, nearly a decade ago, at GW University, since then you've been adjunct faculty as well as a doctoral research advisor at George Washington University. So you probably have some good insights into why people... So let's say a listener out there who's interested in AI, they could be a data scientist, ML engineer, a software developer, maybe a product manager, but amongst our listeners out there, why should someone perhaps consider doing a PhD, maybe even specifically a systems engineering PhD? How would that be helpful to them?

Gina: 00:57:34 I think it was helpful. So I achieved my PhD when I was hired by MITRE in 2011, I believe. Onboarding, they were like, "Yes, we pay for 100% tuition reimbursement."I was like, "Ooh, this is pretty amazing considering I'm sitting on student debt currently." But I was like, "Okay, what am I going to study?" Well, when I entered MITRE, my title was systems engineer, and I thought, "Oh, okay." Then I received a postcard in the mail from GW that says, "We are kicking off our program for PhD in systems engineering." And I was like, "Oh, this is great." I applied, I got in, and here we are. The rest is history.

00:58:27 And so what my advisors did for me, GW asked and said, "Hey, we'd love for you to come in and be an adjunct as well as be a doctoral research advisor." So I've been able to do that and supported 21 students in attaining their doctorates in systems engineering at GW as well. And that's really been super rewarding because not only that

constant learning, I learn from my students because these are folks who are experienced, they're already in the working field and they want to attain their PhD to be able to support them in advancing the mission of their organization or advancing their personal goals. And that's the reason why I got my PhD, is to advance my personal goals as well as support MITRE in some of their mission that they were driving for the federal government.

Jon: 00:59:26 Amazing. When you say personal goals, do you mean that you used your systems engineering education in your personal life to kind of engineer? Is that what you mean?

Gina: 00:59:38 Yes. So I always wanted a PhD, so that was my personal goal.

Jon: 00:59:41 Right. Right.

Gina: 00:59:43 And the universe, I believe in the universe gives you what you need when you need it. I received what I needed at the time at MITRE, and now it's supporting me in AI, in supporting smaller companies and understanding that they have to really lay out their framework. Do they have the right set of skills? Do they have the right developers? Do they have these developers have the skillset to be able to build out their capabilities? Did they do the design? Do they understand their design and what it is that they're building? And what does their roadmap look like in the next three, five years? And who is their market segment that they want to go after? And how do they plan to deploy it? And is ongoing maintenance a part of their lifecycle processes?

01:00:29 So yes, I use it personally for these smaller companies. I use it for myself. I do talks and I present at conferences, and it's helped me to be a really good researcher and to really take on different perspectives to bring together a specific topic.

Jon: 01:00:49 Awesome. All right, well, thank you Gina for taking all this time with us today and providing us with so much fascinating information on a topic that was badly needed to be covered on this show. Gina, before I let you go, do you have a book recommendation for us?

Gina: 01:01:04 You asked me that earlier. I have a 16-year-old daughter who is the only one left at home. I'm going to be an empty-nester in two years, and I'm counting the days. I love my children, but I love that [inaudible 01:01:19] they leave the nest. I have two already left the nest. One is... Oh my goodness, Hannah's birthday is on June 20th. She's going to be 20 years old, not a teenager. She's at Simmons University studying neuroscience and human behavior. Sarah is at University of Pittsburgh, 21 years old, studying biology with a minor in music.

01:01:39 But I read books on raising young women, raising teenagers. Especially during Covid, I was reading that a lot, how to manage, how to support your teens through hard times, through difficult times. I don't have a specific author in mind because I think I read so many, especially around SAT guidance and college preparation guidance and topics like that that were relevant for me and for my family. But those are the books that I've been reading quite a lot. Definitely staying up on podcasts as well as AI research coming out, policies. Yeah, I've been reading those too.

Jon: 01:02:24 All right. Well, thank you for that recommendation, Gina. And yeah, certainly something maybe that we often don't even think about is that in these kinds of challenging personal circumstances or opportunities where there's maybe the possibility of bettering ourselves in some personal way, maybe a lot of us don't often think to pick up a book. We might prefer to stew in concern or talk to a friend or a therapist or that kind of thing. But of course,

books out there could provide a wealth of guidance as well.

Gina: 01:03:00 Absolutely. Oh, yeah. Don't get me wrong. I did go see a therapist, a family therapist for me and my daughter's [inaudible 01:03:11]. I threw them in an RV and we drove across the country and back looking at this beautiful space we call the United States landscape. But yeah, I mean, I read a book on how to do that too.

Jon: 01:03:25 Nice. Awesome, Gina. Well, thank you so much for that personal insight as well. For people who would like to follow you after this episode, what are the best ways that they should do that?

Gina: 01:03:37 They can reach me on LinkedIn. You can provide my email address to them as well, as well as my cell phone number.

Jon: 01:03:43 Wow. Okay. Well, there you go. I've done hundreds of episodes of this show now and nobody has ever offered that before, so I guess we'll have that for you in the show notes. So, wow. Incredibly kind of you, Gina. That's amazing. Fantastic. Thank you for taking all the time today, for being generous with your time and that wildly generous offer to our listeners. We may be catching up with you again soon because I understand you have some exciting developments in the works, which you can't announce yet. But when you can, we'll have you back on to talk about those.

Gina: 01:04:21 I'm really excited to come back and tell you all the cool stuff that I'm working on.

Jon: 01:04:32 What a great episode with Dr. Gina Guillaume-Joseph. In it, Gina filled us in on how 80% of data are unstructured. And of those, 90% are never analyzed. She also talked about how safely digesting these data could vastly boost

the fodder available for training AI models, how responsible AI frameworks like the AI Bill of Rights, the NIST AI risk framework, and the MITRE ATLAS Threat model can be integrated into concrete implementation recommendations for developers of AI driven systems around the world and how systems engineering covers the full lifecycle of development so that projects, including of course AI projects, are delivered on time, on budget, and compliant with regulations.

01:05:14 As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Gina's social media profiles, as well as my own at superdatascience.com/795. Thanks of course to everyone on the Super Data Science team, our Podcast Manager of Ivana Zibert, media editor Mario Pombo, operations manager Natalie Ziajski, researcher Serg Masis, writers Dr. Zara Karschay and Silvia Ogweng, and founder Kirill Eremenko. Thanks to all of them for producing another excellent episode for us today.

01:05:49 For enabling that super team to create this free podcast for you, we are deeply grateful to our sponsors. You can support this show by checking out our sponsors links, which are in the show notes. And if you yourself are interested in sponsoring an episode, you can get the details on how by making your way to jonkrohn.com/podcast. Otherwise, please share with folks that would like this episode. Review this episode on your favorite podcasting app. Subscribe if you're not a subscriber, of course. And yeah, most importantly, I just really hope you'll keep on listening. I'm grateful to have you listening. And I hope I can continue to make episodes you love for years and years to come. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.